1307-4

COMPONENT PART NOTICE

THIS PAPER IS A COMPONENT PART OF THE FOLLOWING COMPILATION REPORT:

TITLE: Computing Science and Statistics: Proceedings of the Symposium on Interface Critical Applications of Scientific Computing (23rd): Biology, Engineering, Medicine, Speech Held in Seattle, Washington on 21-24 April 1991.

TO ORDER THE COMPLETE COMPILATION REPORT, USE AD-A252 938.

— THE COMPONENT PART IS PROVIDED HERE TO ALLOW USERS ACCESS TO INDIVIDUALLY AUTHORED SECTIONS OF PROCEEDING, ANNALS, SYMPOSIA, ETC. HOWEVER, THE COMPONENT SHOULD BE CONSIDERED WITHIN THE CONTEXT OF THE OVERALL COMPILATION REPORT AND NOT AS A STAND-ALONE TECHNICAL REPORT.

THE FOLLOWING COMPONENT PART NUMBERS COMPRISE THE COMPILATION REPORT:

AD#: AD-P007 096 thru AD#: AD-P007 225

AD#: AD#:

AD#: AD#:

Accesion For	
NTIS CRA&I	☑
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and / or Special
A-1	

Exploring Posterior Distributions Using Markov Chains

Luke Tierney*
School of Statistics
University of Minnesota
Minneapolis, MN 55455

Abstract

Several Markov chain-based methods are available for sampling from a posterior distribution. Two important examples are the Gibbs sampler and the Metropolis algorithm. In addition, several strategies are available for constructing hybrid algorithms. This paper outlines some of the strategies that are available, and discusses some theoretical and practical issues in the use of these strategies. In addition, some preliminary efforts to use Markov chains to control dynamic graphics for exploring higher-dimensional posterior distributions are outlined.

1 Introduction

Suppose we are given a posterior distribution π on a quantity θ with values in a space E. Usually E will be a subset of $\mathbb{R}^k$ and π will have a density with respect to a measure μ,

$$\pi(dx) = \pi(x)\mu(dx).$$

For simplicity, π will be used to denote both the distribution and the density. We may be interested in computing a particular numerical characteristic of π, or more generally in developing an understanding of what information π contains about θ.

Several methods for computing characteristics of posterior distributions are now available. These include asymptotic approximations, numerical integration, and sampling or Monte Carlo methods. Sampling methods for examining posterior distributions provide ways of generating samples with the property that the empirical distribution of the sample, or an appropriately weighted empirical distribution, approximate the posterior distribution. Using such samples, it is easy to estimate characteristics such as the mean or standard deviation of a function of θ. Marginal distributions can be estimated using smoothing or, in some cases, variance reduction methods. In addition, for equally weighted samples methods for viewing point clouds, such as rotating plots and Grand Tours, can be used to examine the joint uncertainty about three or more components or features of θ.

A number of different sampling methods are available. In rare cases it is possible to sample directly from the posterior distribution and thus obtain an *i.i.d.* sample from π. In most problems this is not possible. Either the sample has to be dependent, or the distribution used to generate the sample has to be different from π. A method that uses independent samples from a distribution similar to π is importance sampling. The sample is then weighted to make up for the difference between π and the distribution used to generate the sample. Over the past decade, most work on sampling methods for exploring posterior distributions has centered on importance sampling (Geweke, 1989; Stewart, 1979; van Dijk *et al.*, 1978; Zellner and Rossi, 1984; among others). An alternative approach that avoids the need for weights is to use a dependent sample, such as the sample path of a Markov chain.

2 Markov Chain Methods

Markov chain methods generate a sample path from a Markov chain that has π as its stationary distribution. Recent work of Gelfand and Smith (1990) on the Gibbs sampling algorithm has renewed interest in Markov chain methods for exploring posterior distributions. Gelfand and Smith extend the Gibbs sampling algorithm of Geman and Geman (1984), originally developed for Bayesian image reconstruction, to continuous distributions and show how the algorithm can be used in a wide variety of problems.

Markov chain methods have a long history in Mathematical physics dating back to the algorithm of Metropolis *et al.* (1953). The Metropolis algorithm is in fact a general class of algorithms that includes versions of the discrete Gibbs sampler as special cases.

*Research supported in part by grant DMS-9005858 from the National Science Foundation

2.1 The Metropolis Algorithm

Metropolis *et al.* (1953) originally proposed the algorithm now known as the Metropolis algorithm as a method of sampling from the equilibrium distribution of an interacting particle system. The algorithm, which is described in Hammersley and Handscomb (1964, Section 9.3) and Ripley (1987, Section 4.7), was extended by Hastings (1970) and explored further by Peskun (1973).

To define Hastings version of the algorithm, let Q be a Markov transition kernel with

$$Q(x, dy) = q(x, y)\mu(dy).$$

Let $E^+ = \{x : \pi(x) > 0\}$, and assume $Q(x, E^+) = 1$ for $x \notin E^+$. Then define

$$\alpha(x, y) = \min\left\{\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1\right\}$$

for $\pi(x)q(x, y) > 0$. Otherwise, $\alpha(x, y) = 1$. If the Markov chain is currently at $X_n = x$, then the algorithm generates a candidate $Y = y$ for the next state from $Q(x, \cdot)$. With probability $\alpha(x, y)$ this candidate is accepted and the chain moves to $X_{n+1} = y$. Otherwise, the candidate is rejected and the chain remains at $X_{n+1} = x$.

Since

$$\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x),$$

a Metropolis chain with initial distribution π is reversible. Therefore π is an invariant distribution for the chain. Some additional conditions on π and Q are needed to insure that π is also a limiting distribution; these conditions are discussed in Section 3 below. Since the acceptance probability only depends on π through the ratio $\pi(y)/\pi(x)$, the density π only needs to be specified up to a constant of proportionality.

If $q(x, y) = q(y, x)$, *i.e.* q is symmetric, then the acceptance probability $\alpha(x, y)$ simplifies to

$$\alpha(x, y) = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}$$

This is the original form of the algorithm proposed by Metropolis *et al.* (1953). Other forms of the rejection probability are possible, but the form given here can be shown to be optimal within a wide class of possible alternative forms (Peskun, 1973).

The Metropolis algorithm is actually a class of algorithms. Each different choice of the kernel Q for generating candidate steps produces a different version of the algorithm. Several classes of kernels appear to be particularly useful for examining posterior distributions.

2.1.1 Random Walk Chains

For $E = \mathbb{R}^k$ and f a density on E, set $Y = x + Z$, with Z drawn independently from f. Then

$$q(x, y) = f(y - x).$$

Thus the kernel Q driving the Metrolopis chain is a random walk. Natural choices of f are normal, uniform, and t distributions. Split-t distributions (Geweke, 1989) may also be useful. The scale matrix for f can be taken as a constant c times the inverse information at the posterior mode. Good choices for the step size constant c are still an open problem, but $c = 1$ and $c = 1/2$ seem to work reasonably well in a number of examples.

If f is symmetric about the origin, *i.e.* if $f(z) = f(-z)$, then q is symmetric and the simpler form of the acceptance probability $\alpha(x, y)$ can be used.

2.1.2 Independence Chains

Suppose f is a density on E, and we generate candidates Y independently from the single density f. Then

$$q(x, y) = f(y).$$

The chain of candidates driving this Metropolis chain is an *i.i.d.* sequence from the density f. The acceptance probability for an independence chain can be written as

$$\alpha(x, y) = \min\left\{\frac{w(y)}{w(x)}, 1\right\}$$

for $w(x) = \pi(x)/f(x)$. The function w is the weight function that would be used in importance sampling when the sample is generated from the density f.

There are a number of similarities between an independence chain and the corresponding importance sampling process. While an independence chain does not require explicit use of the weights, it will rarely accept candidates with low weights. On the other hand, a candidate with high weight will almost always be accepted. Furthermore, when the chain reaches a point x with high weight $w(x)$, it will usually remain there for several iterations, thus building up weight on x within the sample path by repetition. Another similarity to importance sampling is that the sample sequence is closer to an *i.i.d.* sequence from π the closer the weight function w is to a constant.

Because of these similarities to importance sampling, it is reasonable to conjecture that guidelines developed for choosing importance sampling densities also apply to choosing densities for driving independence chains. In particular, it is advisable to choose a density with thicker tails than π and thus a bounded weight function

w. Families like the split-t that produce good importance sampling densities are likely to be good choices for independence chains.

2.1.3 Rejection Sampling Chains

An interesting special case of an independence chain occurs when the density f is sampled using rejection sampling. In attempting to use rejection sampling to sample directly from π, we use a density h and a constant c such that, hopefully, $\pi(x) < ch(x)$. If we repeat the process of sampling Z from h and then U uniformly from $[0, ch(Z)]$, until $U < \pi(Z)$, then the final value of Z has density

$$f(x) \propto \pi(x) \wedge ch(x).$$

If we do indeed have $\pi(x) \leq ch(x)$, then f is proportional to π and we obtain an *i.i.d.* sample from π. But it is very difficult to insure that c is large enough for ch to dominate π without choosing c excessively large, leading to an inefficient algorithm with many rejections. And even then without extensive analysis of the tails of h and π we cannot be certain that ch does dominate π.

Fortunately, using this rejection scheme to drive an independence Metropolis chain provides a simple remedy. If we do have $\pi(x) \leq ch(x)$ for all x, then the weight function w is a constant, no candidates are rejected, and the rejection process produces an *i.i.d.* sequence from π that is simply passed through the Metropolis algorithm unchanged. But if ch does not dominate π for some x, then, when the chain reaches such an x, the Metropolis algorithm will occasionally reject candidate steps in order to build up mass on this x to make up for the deficiency in the envelope ch. This introduces some dependence, but insures that the equilibrium distribution of the sample path is π even if the envelope is deficient.

2.2 Combining Strategies

The Gibbs sampler and the Metropolis algorithms described above provide a number of Markov chain strategies. In addition to choosing any one of these strategies and using it in its pure form, it is possible to form hybrid strategies.

Suppose $P_1, \ldots, P_m$ are Markov kernels with invariant distribution π. Two simple ways of combining these kernels is as a mixture or a cycle. In a mixture, probabilities $\alpha_1, \ldots, \alpha_m$ are specified, and at each step one of the kernels is selected according to these probabilities. In a cycle, each kernel is used in turn, and when the last one is used the cycle is restarted.

Both strategies can be used in several ways. For example, a Gibbs sampler can be combined with occasional steps from an independence chain in a mixture or a cycle to "restart" the Gibbs sampler and thus reduce correlations while preserving the equilibrium distribution. As another example, suppose θ can be split into two components (θ_1, θ_2), and direct sampling from $\theta_1|\theta_2$ is possible but direct sampling from $\theta_2|\theta_1$ is not possible. Such a situation is considered by Zeger and Karim (1991). Then "Gibbs steps" for $\theta_1|\theta_2$ can be combined with Metropolis steps for $\theta_2|\theta_1$ in a mixture or a cycle.

3 Some Theoretical Results

Whatever approach is used to produce a Markov chain with invariant distribution π, before the chain can be used with confidence to generate samples for examining π certain theoretical questions need to be addressed. Answers to some of these questions can be obtained using some recent developments in general state space Markov chain theory as described, for example, in Nummelin (1984). This section outlines this approach. A more complete discussion is given in Tierney (1991).

3.1 Convergence

The first question to be addressed is whether the invariant distribution π is also the equilibrium distribution for the chain, *i.e.* whether the distribution of the chain after n iterations converges to π. In discrete state space Markov chain theory, two conditions are needed: irreducibility and aperiodicity. The same is true in general state space theory. Periodicity for general state spaces can be defined in much the same way as for discrete spaces. The concept of irreducibility is a little more complicated, since individual states are usually not hit with positive probability. It is therefore necessary to speak of irreducibility with respect to a measure. In the present context, a natural choice for this measure is π itself. We will therefore say that a Markov chain is π-irreducible if for every set A with $\pi(A) > 0$ the probability of the chain ever entering A is positive for every starting point x of the chain.

Irreducibility and aperiodicity need to be verified for each Markov chain. Some useful sufficient conditions are available for certain Metropolis chains. For example, a random walk chain is π-irreducible and aperiodic if the increment density is positive on a neighborhood of the origin and the density π is positive on all of $\mathbb{R}^k$. An independence chain is π-irreducible and aperiodic if the candidate generation density f is positive whenever the density π is positive.

If a chain with invariant distribution π-irreducible and aperiodic, then it can be shown that the chain must be

positive recurrent and that for π-almost all x,

$$\| P^n(x,\cdot) - \pi(\cdot) \| \to 0$$

where $\| \cdot \|$ denotes the total variation distance and P^n is the distribution after n steps of the chain started at x.

If the chain is *Harris recurrent*, then this convergence occurs for all x. The definition of Harris recurrence is somewhat technical, but a simple sufficient condition is available that is satisfied by all π-irreducible Metropolis chains and essentially all π-irreducible Gibbs samplers.

A π-irreducible aperiodic Markov chain with invariant distribution π is called *ergodic* if it is aperiodic and positive Harris recurrent.

3.2 Rates of Convergence

Once we know that the distribution of a chain converges to π, the next question is to determine the rate of convergence. The theory presented in Nummelin (1984) provides several classifications for rates of convergence of ergodic chains:

Degree 2: If a chain is ergodic of degree 2, then

$$n \| P^n(x,\cdot) - \pi(\cdot) \| \to 0$$

for π-almost all x.

Geometric: An ergodic chain is geometrically ergodic if $\| P^n(x,\cdot) - \pi(\cdot) \| \leq M(x) r^n$ for some $r < 1$ and some function M with $\int M d\pi < \infty$.

Uniform: An ergodic chain is called uniformly ergodic if $\| P^n(x,\cdot) - \pi(\cdot) \| \leq M r^n$ for some $r < 1$ and some constant M.

Uniform ergodicity is the strongest of these forms of convergence and it is the easiest form to work with. A necessary and sufficient condition for a chain with kernel P to be uniformly ergodic is that there exist a probability ν, a constant $\beta > 0$ and an integer $n \geq 1$ such that $\nu(A) \leq P^n(x,A)$ for all A and x. Using this condition, it is possible to derive a variety of sufficient conditions for uniform ergodicity For example, if $\mu(E) < \infty$ and the densities q and π are bounded and bounded away from zero, then the corresponding Metropolis kernel is uniformly ergodic. As another example, an independence Metropolis kernel is uniformly ergodic if the weight function $w(x)$ is bounded.

This condition can also be used to derive conditions for uniform ergodicity of hybrid kernels in terms of conditions on the component kernels. For mixtures the condition is particularly simple: if P is uniformly ergodic, then any mixture using P with positive probability is uniformly ergodic. For cycles a slightly more complicated condition appears to be needed: if P is used in a cycle and there exists a probability ν and a constant $\beta > 0$ such that $\nu(A) \leq P(x,A)$ for all A and x, then the cycle is uniformly ergodic. This condition is satisfied if P is an independence kernel with a bounded weight function. Combining such a kernel in a mixture or a cycle with any other kernel, such as a Gibbs kernel, therefore insures that the hybrid chain is uniformly ergodic. This provides theoretical support for using occasional independence "restart" steps together with a Gibbs sampler to improve the properties of the sampler.

3.3 Limiting Behavior of Averages

In Markov chain methods, sample path averages are used to estimate expectations under the distribution π. A law of large numbers and a central limit theorem insure that these estimates converge at reasonable rates. The law of large numbers follows from the ergodic theorem and needs no conditions other than existence of the expectation under π:

Law of Large Numbers. *If P is ergodic with invariant distribution π, and $\pi|f| < \infty$, then for any initial distribution*

$$\bar{f}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \pi f = \int f(x)\pi(dx)$$

almost surely.

A central limit theorem does appear to require some conditions on the rate of convergence of the chain:

Central Limit Theorem. *If P is ergodic of degree 2 with $\pi P = \pi$, and f is bounded, then for any initial distribution the distribution of*

$$\sqrt{n}(\bar{f}_n - \pi f)$$

converges weakly to a normal distribution with mean zero and variance $\sigma^2(f)$.

Weaker but more complicated sufficient conditions are available. Expressions for the asymptotic variance $\sigma^2(f)$ are available for finite E (Peskun, 1973; Kemeny and Snell, 1976). Other expressions involving certain hitting times are available for general state spaces (Nummelin, 1984). These expressions do not appear to be useful for computing the asymptotic variance.

4 Using a Markov Chain

Once a Markov chain strategy with satisfactory theoretical properties has been selected, it can be used to estimate numerical characteristics or to provide graphical views of features of the posterior distribution.

4.1 Numerical Uses

Using Markov chains for calculating numerical characteristics of a posterior distribution is in principle straight forward: expectations with respect to π can be approximated by sample path averages. There are, however, a number of issues that need to be considered before running a chain.

4.1.1 Choosing a Sampling Plan

The first issue concerns the choice of a sampling plan. There are two extreme approaches. Several authors have proposed that Markov chains should be used to generate n independent realizations from the posterior by using n separate runs, each of length m, and retaining the final states from each chain. The run length m is to be chosen large enough to insure that the chain has reached equilibrium. An alternate approach is to use a single long run, or perhaps a small number of long runs. Experience and theoretical assessments in the simulation literature appear to favor the use of long runs (Bratley *et al.*, 1987, Section 3.1.1; Kelton and Law, 1984). The major drawback of using short runs is that it is virtually impossible to tell when a run is long enough based on such runs. Even using long runs, determining how much of the initial series is affected by the starting state is very difficult, but some literature on the subject is available (Ripley, 1987, Section 6.1). A second drawback of short runs is that it makes inefficient use of the data: only n out of a total of nm data points are used. With a single run of length nm it is possible to use all the data, after possibly discarding a small initial fraction.

A complication that does arise from the dependence in using a single series is that variances of estimates are hard to assess. Again the simulation literature offers several alternatives, such as the use of batch means and time series analysis (Bratley *et al.*, 1987, Chapter 3; Ripley, 1987, Chapter 6). For some purposes it may nevertheless be useful to have an approximate independent sample from the posterior. Using long runs this can be achieved by retaining every r-th point of a sample path. The number r of points to skip in order to produce approximate independence can usually be chosen much smaller than the number m of steps needed to reach approximate equilibrium, since small amounts of correlation are usually much less serious than biases in estimates of means.

4.1.2 Determining the Run Length

Another consideration is to determine the total sample size or run length required for accurate estimates. For an *i.i.d.* sample of size n, the standard deviation of the sample mean of a function $f(\theta)$ is $\sigma/\sqrt{n}$, where σ is the posterior standard deviation of $f(\theta)$. If a preliminary estimate of σ is available, perhaps from an asymptotic analysis, then this can be used to estimate the sample size that would be required in *i.i.d.* sampling. In dependent sampling, observations are generally positively correlated and a larger simple size will be required. If the series is modeled as a first order autoregressive process, then the standard deviation of the sample mean is

$$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{1+\rho}{1-\rho}}$$

where again σ is the posterior standard deviation of $f(\theta)$ and ρ is the autocorrelation of the series $f(X_n)$. A rough estimate of ρ can thus be used to adjust the sample size for dependence in the series.

Instead of determining a fixed sample size in advance, it is also possible to use sequential or batch sequential rules for determining when to stop sampling. Since prior information on the values of the posterior mean and standard deviation is often available form initial analyses, Bayesian sequential methods are a natural choice. Batching can be used to insure that an assumption of normality for batched means is reasonable.

One sequential approach that should be avoided is to plot successive sample means and stop sampling when the means appear to have converged. Since sample means change by increments on the order of $O(n^{-1})$ but errors are of order $O(n^{-1/2})$, this approach will produce sample sizes that are too small. The presence of positive correlations in Markov chain series makes these series appear to have converged even earlier, even though the correlations imply that errors are larger and thus larger sample sizes are required than with *i.i.d.* sampling.

4.1.3 Numerical Issues

Some consideration of numerical stability is needed in using any sampling based method. Expressions used to evaluate log posterior densities obtained by translating mathematical formulas into a computer language are often reasonably stable near the posterior mode but not far away from the posterior mode. This can lead to overflows or, on IEEE hardware, results that are NAN's or INF's. One way to avoid these problems is to carefully

study the formulas for evaluating the log posterior density and modify them to be numerically stable even for extreme parameter values. The effort required to do this can be considerable. An expedient alternative that is often effective is to truncate the parameter space to a reasonable range that contains essentially all the posterior probability and for which the posterior density formula is numerically stable. This truncation also often insures that a Markov chain used to sample from π is uniformly ergodic and thus improves the behavior of the Markov chain estimates.

The need to allow truncation is an important consideration in developing software for implementing sampling based methods. Subroutines must allow for user supplied range test functions or allow the results returned by the log posterior subroutine to indicate a parameter that is outside of the range.

A numerical issue that is unique to Markov chain methods is the possibility that rounding may introduce absorbing states. If this happens, results obtained from a Markov chain method may be meaningless. Again truncation away from areas of the state space where such rounding may occur can be helpful.

4.1.4 Variance Reduction

As with any simulation method, variance reduction techniques can often significantly reduce the sample sizes required for accurate estimates. Standard variance reduction methods such as importance sampling, antithetic variates, conditioning, and control variates (Bratley *et al.*, 1987, Chapter 2; Ripley, 1987, Chapter 5) can be used with any Markov chain method.

Importance sampling can be used as a variance reduction method by using a Markov chain with equilibrium distribution f instead of π and then weighting sample results with appropriate importance weights. Conditioning is often useful in Gibbs samplers, since the assumptions required for the Gibbs sampler imply that conditional means or densities of one parameter given the rest are usually available. Gelfand and Smith (1990) refer to this use of conditioning as Rao-Blackwellization.

Antithetic variation can be introduced into a Markov chain method by using a Metropolis step in which a candidate step is obtained by reflecting the current state of the chain through a point. If the posterior density is approximately symmetric about this point, then the sample will be also, and the resulting negative correlations will reduce variances of estimates of linear functions of θ. This technique can also be used to take advantage of approximate axial symmetries in a posterior distribution.

One way to introduce control variates into a Markov chain method is to use the sample path with importance weights to calculate estimates of normal approximations and to correct for the errors in these estimates.

4.1.5 Monitoring Sampler Performance

In using Markov chain methods, it is important to monitor the performance of the samplers to insure that they are not exhibiting any unusual behavior. Gelfand and Smith (1990) propose the use of quantile plots to monitor performance. Monitoring sample paths of estimates is also useful for this purpose, as is monitoring autocorrelations of the parameters. Adaptive time series models may also be useful for determining whether a series exhibits any unusual features.

For Metropolis chains it is also important to keep track of the the number of candidates that are rejected. For an independence chain, the proportion of rejections can be related to the total variation distance between the posterior density π and the candidate generation density f.

By monitoring the performance of a sampler, in particular in the early stages, it is possible to experiment with different setting for sampler parameters to obtain samplers that are efficient for a particular problem. More work is needed to find good strategies for making such parameter adjustments.

4.2 Graphical Uses

Numerical summaries, such as posterior means, standard deviations, marginal densities, and correlations, provide insight into the uncertainty about one or perhaps two features of θ at a time. For understanding uncertainty in higher dimensions graphical methods may be more useful than numerical summaries.

4.2.1 Plotting Samples

For three-dimensional quantities, one useful graphical method available on microcomputers and workstations with bitmapped displays is a rotatable three-dimensional scatterplot. By selecting every r-th entry in a Markov chain sample path we can obtain an approximate *i.i.d.* sample from the posterior distribution and display this sample in a rotatable scatterplot. Three-dimensional structures will readily become apparent as the point cloud of the sample is rotated.

Rotatable scatterplots are only useful for examining three dimensions at a time. A method that may be useful for higher dimensions is the Grand Tour. Again an approximate *i.i.d.* sample can be selected and displayed in a Grand Tour. Implementations of the Grand Tour

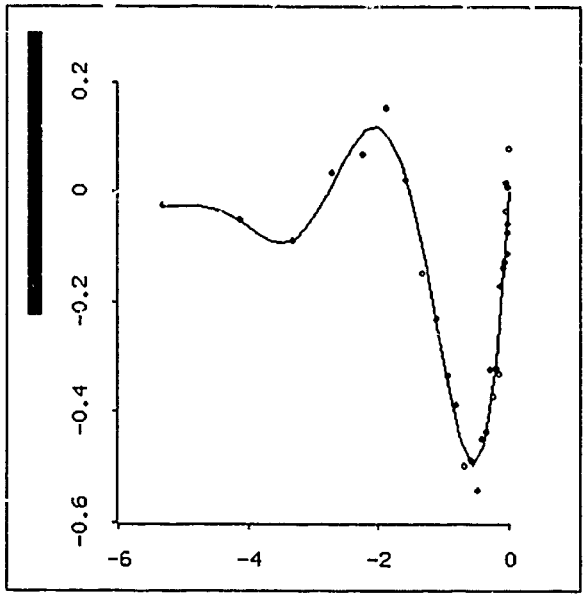

Figure 1: Posterior mean of a response function.

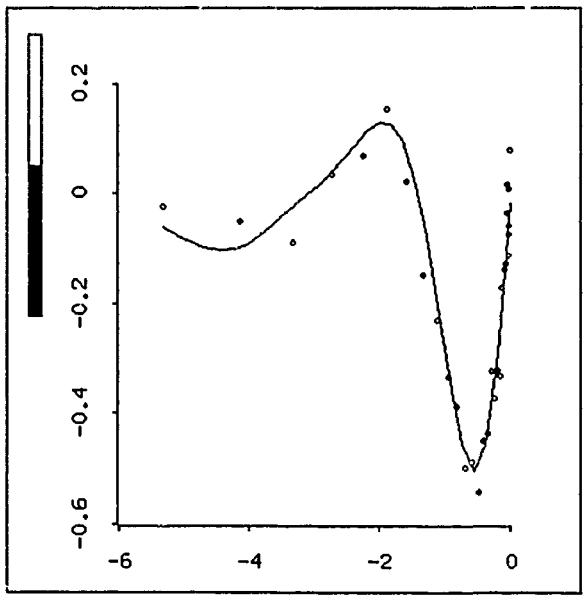

Figure 2: A second response function supported by the posterior distribution.

are only now becoming widely available, so extensive experience with this method is not yet available. Early results suggest that this method is reasonably effective for detecting structures in four to six dimensions.

4.2.2 Controlling Animations

If θ is more than five- or six-dimensional, then it may be difficult enough to understand θ itself, much less uncertainty about θ. If a graphical view of θ is available that is meaningful for particular values of θ, then one way of developing an understanding of the uncertainty about θ is to look at an animated version of the graph in which θ is moved through a variety of values that are plausible under the posterior distribution.

As an example, suppose we have a smooth response function θ of a real variable x in some interval I that is measured with error. Thus we obtain measurements of the form

$$Y = \theta(x) + \varepsilon.$$

Our prior opinion on the function θ suggests that this function is smooth, but does not suggest any particular parametric structure.

Several approaches are available for specifying such a prior distribution. Most involve choosing a prior on coefficients in some representation, such as a power series or spline. The coefficients of these representations are not likely to be particularly meaningful. But a plot of the response function θ over the interval I is readily understood. Figure 1 shows a plot of the posterior mean of θ for a particular example. This mean exhibits a number of features, such as a pronounced global minimum and a secondary local minimum. Are these features really present in θ or are they merely artifacts of the posterior mean? One way to answer this question is to look at other functions θ that are supported by the posterior distribution. This can be done by running an animation that shows graphs of different values of θ.

To provide a good understanding of the posterior distribution, an animation needs to visit all areas supported by the posterior. In addition, to allow the user to keep track of the changes in θ as it moves through the posterior distribution, the animation has to move smoothly. These objectives can be achieved using a random walk-driven Metropolis chain with the posterior distribution as its equilibrium distribution. Using the posterior as the equilibrium insures that the chain does eventually approach all possible values of θ but spends most of its time near values that are better supported by the posterior distribution. The correlation in the random walk insures that the chain moves in small steps, thus providing the visual continuity that is necessary for an effective animation. Thus the correlations in the Metropolis chain that are a nuisance for numerical computations are in fact an advantage for this graphical application. Continuity can be further enhanced by interpolating between steps of the random walk.

Figure 2 shows another view of the animation. Viewing the animation for this particular example for a few

minutes quickly reveals that the global minimum is quite well defined but the shape of the left half of the curve is very uncertain.

A useful enhancement for this animation is the bar shown at the left of the two plots. The solid part of the bar represents the probability content in the posterior at or below the level of the current θ, computed using a χ^2 approximation. This gives a quick indication of how plausible the current view is.

Many variations on this animation are possible. For example, using the posterior distribution as the equilibrium of the driving Markov chain is a reasonable starting point but is not essential. At times it may be useful to force the chain to concentrate its motion closer to the mode, or to move farther away from the mode and possibly find interesting features that are farther away. This can be accomplished by using a Markov chain with an equilibrium density that is a power of the posterior density – by "cooling" or "heating" the posterior distribution in the terminology of simulated annealing.

Much additional work is needed to explore ways of merging numerical methods such as the ones described in this paper with new computing hardware that is now becoming more widely available. The animation described here is a first step in that direction.

References

Bratley, P., B. L. Fox, and L. E. Schrage (1987). *A Guide to Simulation.* New York, NY: Springer, second edition.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **6** 721–741.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57** 1317–1339.

Hammersley, J. M. and D. C. Handscomb (1964). *Monte Carlo Methods.* London: Chapman and Hall.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

Kelton, D. W. and A. M. Law (1984). An analytical evaluation of alternative strategies in steady-state simulation. *Operations Research* **32** 169–184.

Kemeny, J. G. and J. L. Snell (1976). *Finite Markov Chains.* New York, NY: Springer.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *J. Chemical Physics* **21** 1087–1091.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators.* Cambridge: Cambridge University Press.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60** 607–612.

Ripley, B. D. (1987). *Stochastic Simulation.* New York, NY: Wiley.

Stewart, L. T. (1979). Multiparameter univariate Bayesian inference. *J. Amer. Statist. Assoc.* **74** 684–693.

Tierney, L. (1991). Markov chains for exploring posterior distributions. Technical Report 560, School of Statistics, University of Minnesota.

van Dijk, H. K., J. P. Hop, and A. S. Louter (1978). An algorithm for the computation of posterior moments and densities using simple importance sampling. *The Statistician* **36** 83–90.

Zeger, S. L. and M. R. Karim (1991). Generalized linear models with random effects: A Gibb's sampling approach. *J. Amer. Statist. Assoc.* **86** 79–86.

Zellner, A. and P. E. Rossi (1984). Bayesian analysis of dichotomous quantal response models. *J. of Econometrics* **25** 365–393.